Shapley and Banzhaf Vectors of a Formal Concept

Dmitry I. Ignatov[2,3] and Léonard Kwuida[1]

[1] Bern University of Applied Sciences, Bern, Switzerland
leonard.kwuida@bfh.ch
[2] National Research University Higher School of Economics, Moscow, Russia
dignatov@hse.ru
[3] St. Petersburg Department of Steklov Mathematical Institute of Russian Academy of Sciences, Russia

Abstract. We propose the usage of two power indices from cooperative game theory and public choice theory for ranking attributes of closed sets, namely intents of formal concepts (or closed itemsets). The introduced indices are related to extensional concept stability and based on counting generators, especially those that contain a selected attribute. The introduction of such indices is motivated by the so-called interpretable machine learning, which supposes that we do not only have the class membership decision of a trained model for a particular object, but also a set of attributes (in the form of JSM-hypotheses or other patterns) along with individual importance of their single attributes (or more complex constituent elements). We characterise computation of Shapley and Banzhaf values of a formal concept in terms of minimal generators and their order filters, provide the reader with their properties important for computation purposes, and show experimental results.

Keywords: Shapley value, Banzhaf value, Interpretable Machine Learning, formal concepts, closed itemsets

1 Introduction

Concept stability indices were introduced to assess robustness of JSM-hypotheses in [13] under deletion of subsets of objects. JSM-method is known as a logical rule-based classification method named after John Stuart Mill [5], which had was later formulated in terms of Formal Concept Analysis (FCA) [14]. In FCA terms, each JSM-rule is an implication of the form "concept intent" $\rightarrow$ "target attribute", where the target attribute corresponds to a predicted class[4].

Concept stability indices allow to rank those intents (or hypotheses) by their robustness under objects deletion and provide the evidence of their non-random nature similarly to bootstrap estimation [15] and swap-permutations [7]. Later on, the notion of stability was rediscovered under the name of robustness of

[4] Usually, each left-hand side includes only a minimal intent with $minsup = 2$ that is not included in the intents of examples of other classes.

 Published in Francisco J. Valverde-Albacete, Martin Trnecka (Eds.): Proceedings of the 15th International Conference on Concept Lattices and Their Applications, CLA 2020, pp. 259–271, 2020.

closed itemsets [23]. On the one hand the authors of the robustness of itemsets considered a more general case, namely, they covered itemsets, closed itemsets (concept intents), and free itemsets (intent generators), but on the other hand they even assumed Bernoullian character of object deletion with a fixed probability α. In [11], by means of Möbius function, it is shown that robustness and stability are analytically equivalent for $\alpha = 0.5$.

However, JSM-hypotheses are not atomic patterns, and even if we know that a certain hypothesis is stable, we cannot judge the importance of each attribute for making the hypothesis stable. To fill the gap, we addressed the Shapley value importance of attributes from Interpretable Machine Learning (IML) [19]. The Shapley value, a notion from Collaborative Game Theory [21], is used to provide a fair payoff to players, which are actually the attributes of an object under classification in our case. By means of the payoff value we can rank the attributes and judge what the contribution or importance of each attribute is for the classification of this object to the predicted class. Being model-agnostic, however this technique assumes a probabilistic output $p(class|object)$ of a classifier and each attribute's payoff is computed as conditional expectation of a related value function over all the subsets of attributes with and without a selected attribute.

Here, we assume that attributes are players of the game where the (extent) stability of a concept is shared between them using the classic Shapley value formula. It has several important properties: efficiency, symmetry, linearity, null or dummy player. For example, efficiency means the sum of the Shapley values of all players equals the value of their coalition, so that the total payoff is distributed among the agents. Another variants of power (importance) indices from Coalition Game Theory are possible, for example, the Banzhaf index [2], though without fulfilling all the mentioned properties.

In this paper, we would like to explain how the Shapley values of individual attributes in a formal concept intent can be computed by means of a Boolean valuation function related to extent stability of the considered formal concept. Moreover, we would like to characterise the Shapley vector of a formal concept in order-theoretic terms and find its relationship with stability indices.

The paper is organised as follows. In Section 2, related work on interpretable machine learning and Shapley values from Game Theory is summarised. Section 3 recalls basics of FCA and concept stability indices. Section 4 illustrates how to use Shapley values to estimate the importance of separate attributes related to stability indices, and introduces the weak Banzhaf index. Section 5 is devoted to machine experiments with model and real machine learning data.

2 Shapley value in IML and FCA communities

2.1 Interpretable Machine Learning

In early 90's, the discipline of Data Mining emerged as a step of Knowledge Discovery in Databases (KDD) process, which was defined as follows: "KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately

understandable patterns in data" [4]. Recently, machine learning researchers realised the necessity of interpretation for a wide variety of black-box models and even for ensemble rule-based methods when many attributes are involved [16][5].

The author of a book [19] on interpretable machine learning notes that there is no mathematical definition of interpretability. In [8] interpretability is understood as "the degree to which a human can consistently predict the model's results". Thus, machine learning models can be ranked according to their interpretability: "A model is better interpretable than another model if its decisions are easier for a human to comprehend than decisions from the other model" [19].

Among the family of approaches, [19] poses global vs. local interpretations. Thus, the global interpretability "is about understanding how the model makes decisions, based on a holistic view of its features and each of the learned components such as weights, other parameters, and structures", while the local interpretability is focused on "a single instance and examine what the model predicts for this input, and explain why" [19]. From this point of view, JSM-rules provides local interpretations when we deal with a particular classification example.

The Shapley value attribution is a model-agnostic method and produces ranking of individual attributes by their importance for classification of a particular example, i.e. it provides a decision maker with local interpretations. Indeed, this attribution methodology is equivalent to the Shapely value solution to value distribution in Cooperative Game Theory [21]. Strumbelj and Kononenko [22] consistently show that the Shapley value is the only solution for the problem of single attribute importance (or attribution problem) measured via the difference between model's prediction with and without this particular attribute across all possible subsets of attributes.

Lundberg and Lee [17] extend this approach further under the name SHAP (Shapley Additive explanation) values [17]. To compute SHAP value for an example x and an attribute m the authors define $f_x(S)$, the expected value of the model prediction conditioned on a subset S of the input attributes, which can be approximated by integrating over samples from the training dataset.

SHAP values combine these conditional expectations with the classic Shapley values from game theory to assign ϕ_m value to each attribute:

$$\phi_m = \sum_{S \subseteq M \setminus \{m\}} \frac{|S|!(|M| - |S| - 1)!}{|M|!} \left(f_x(S \cup \{m\}) - f_x(S) \right), \tag{1}$$

where M is the set of all input attributes, S is a certain coalition of players, i.e. subset of attributes $S \subseteq M$.

In our study, we will follow classical definition of the Shapley value [21], where a monotone Boolean function is used instead of expected value $f_x(S)$. Thus, we evaluate the contribution of each attribute with respect to extent stability rather than find its importance for a certain example classified by JSM-hypotheses.

[5] The workshops on Interpretable Machine Learning: `https://sites.google.com/view/whi2018` and `https://sites.google.com/view/hill2019`

2.2 Shapley value in FCA community

In [3], the authors introduced cooperative games on concept lattices. Any game of this type induces a game on the set of objects, and a game on the set of attributes. In these games the notion of Shapley value naturally arises as a rational solution for distributing the total worth of the cooperation among the players. The authors of [18] studied algorithms to compute the Shapley value for a cooperative game on a lattice of closed sets given by an implicational system; the main computational advantage of the proposed algorithms is based on maximal chains and product of chains of fixed length.

3 FCA basics and concept stability indices

Let $\mathbb{K} := (G, M, I)$ be a formal context; that is a triple of sets such that $I \subseteq G \times M$. We call G the set of objects and M the set of attributes. For $A \subseteq G$ and $B \subseteq M$, we define the derivation operation by:

$$A' := \{m \in M \mid \forall a \in A : (a, m) \in I\} \text{ and } B' := \{g \in G \mid \forall b \in B : (g, b) \in I\}.$$

A pair (A, B) is called a formal concept of $\mathbb{K}$ if $A' = B$ and $B' = A$. In that case, A is called the *extent* and B the *intent* of the concept (A, B). Let (A, B) be a formal concept of $\mathbb{K}$ with $|A| = l$, and $|B| = n$. To define the stability indices as in [13,10,20,12], we first recall the notion of generator.

Definition 1. *A generator of the concept intent B is any $Y \subseteq B$ such that $Y'' = B$. A set $X \subseteq B$ is a minimal generator of B iff $X'' = B$ and no proper subset of X is a generator of B. The sets of all generators (resp. all minimal generators) of B will be denoted by* $\mathrm{gen}(B)$ *(resp.* $\mathrm{mingen}(B)$*).*

Definition 2. *In [10], the stability index $J_k(A, B)$ (or simply $J_k(B)$) of the k-th level is defined as follows:*

$$J_k(A, B) := \frac{|\{Z \subseteq A,\ |Z| = k \text{ and } Z' = B\}|}{\binom{l}{k}} = \frac{|\{Z \subseteq A,\ |Z| = k \text{ and } Z'' = A\}|}{\binom{l}{k}}.$$

Extensional and intensional stability indices $\sigma_i(A, B)$ and $\sigma_e(A, B)$ were introduced in [20,12], as variations of integral stability from [10]. Here the proportions are taken over all subsets, instead of subsets of size k.

Definition 3. *The intensional stability index of a concept (A, B) is defined by:*

$$\sigma_i(A, B) := \frac{|\{Z \subseteq A \text{ such that } Z' = B\}|}{2^l} = \frac{|\{Z \subseteq A \mid Z'' = A\}|}{2^l} = \frac{|\mathrm{gen}(A)|}{2^l}.$$

Similarly the extensional stability index of a concept (A, B) is defined by:

$$\sigma_e(A, B) := \frac{|\{Y \subseteq B \text{ such that } Y' = A\}|}{2^n} = \frac{|\{Y \subseteq B \mid Y'' = B\}|}{2^n} = \frac{|\mathrm{gen}(B)|}{2^n}.$$

We call $J_k(C)$ the intensional stability index of the k-th level, while the extensional stability index of the k-th level is defined similarly.

Definition 4. *The extensional stability index of the k-th level:*

$$J_k(A) = |\{Y \subseteq B \mid |Y| = k, Y' = A\}| / \binom{n}{k}.$$

4 Shapley values for attribute importance in terms of concept stability indices

Let us consider the extent stability index $\sigma_e(A, B)$ as in Definition 3. Its numerator induces a monotone Boolean function $v_B : 2^B \to \{0, 1\}$, which may play a role of the value function of any $Y \subseteq B$ considered as a coalition of players, i.e. attributes in our case:

$$v(Y) = \begin{cases} 1, & \text{if } Y' = A \text{ and } Y \neq \emptyset \\ 0, & \text{otherwise} \end{cases}.$$

We will omit B as a lower index of v if it does not result in notation collisions. Note that $\sigma_e(A, B) = \Big(\sum\limits_{Y \subseteq B} v(Y) + [\emptyset' = B]\Big)/2^{|B|}$, where $[P]$ is the Iverson notation for the value of a predicate P[6].

Definition 5. *The Shapley value of $m \in B$ for a concept (A, B) is defined by:*

$$\varphi_m(A, B) = \frac{1}{|B|} \sum_{Y \subseteq B \setminus \{m\}} \frac{1}{\binom{|B|-1}{|Y|}} \big(v(Y \cup \{m\}) - v(Y)\big).$$

The vector of Shapley values for all the attributes from B is called the (reduced) Shapley vector of (A, B), and its extension to all attributes ($\varphi_m(A, B) = 0$ for $m \notin B$) is called the extended Shapley vector of (A, B).

Since minimal generators are important coalitions and the function $v(\cdot)$ is inspired by the stability indices, we would like to study their interconnection. The Shapley values of attributes in case of extent stability can be expressed with the help of only minimal generators.

Theorem 1. *Let $m \in X_m \in \mathrm{mingen}(B)$ and $m \notin X_{\overline{m}} \in \mathrm{mingen}(B)$.*[7] *Then*

$$\varphi_m(A, B) = \frac{1}{|B|} \sum_{D \sqcup \{m\} \in \bigcup X_m\uparrow \setminus \bigcup X_{\overline{m}}\uparrow} \frac{1}{\binom{|B|-1}{|D|}}.$$

Proof. Let us consider the three possible cases of the expression under summation.

[6] $[P] = 1$ if P is true and $[P] = 0$ otherwise

[7] $\sqcup$ is disjoint union, X_m and $X_{\overline{m}}$ are minimal generators of B with and without m, respectively.

$v(D \cup \{m\})$	$v(D)$	$v(D \cup \{m\}) - v(D)$
0	0	0
1	0	1
1	1	0

1. When D and $D \cup \{m\}$ are not generators of B, we have the first case (first row of the table); they do not contribute to $\varphi_m(A, B)$. In this case those sets are generators of less general concepts with intents included in B.
2. When $D \notin \text{gen}(B)$ and $D \cup \{m\} \in \text{gen}(B)$ hold, then we have second case. By definition each generator of B is in an interval formed by a minimal generator and B, i.e. $X\uparrow = [X, B]$, where $X \in \text{mingen}(B)$. We are interested in all $S \sqcup \{m\} \in \text{gen}(B)$ such that $S \notin \text{gen}(B)$. If $S \notin \text{gen}(B)$, then $(S', S'') > (A, B)$. Since all $D \in [S, S'']$ are not generators of B, then $v(D) = 0$. On the other hand, $S \sqcup \{m\} \subseteq D \sqcup \{m\}$, then $D \sqcup \{m\} \in \text{gen}(B)$ and $v(D \sqcup \{m\}) = 1$.

If $S \sqcup \{m\} \in X_{\overline{m}}\uparrow$, then $X_{\overline{m}} \subseteq S$ and $v(S) = 1$. Thus we need not to consider all $S \sqcup \{m\}$ from $\bigcup X_{\overline{m}} \uparrow$, the set of order ideals of minimal generators not containing m.

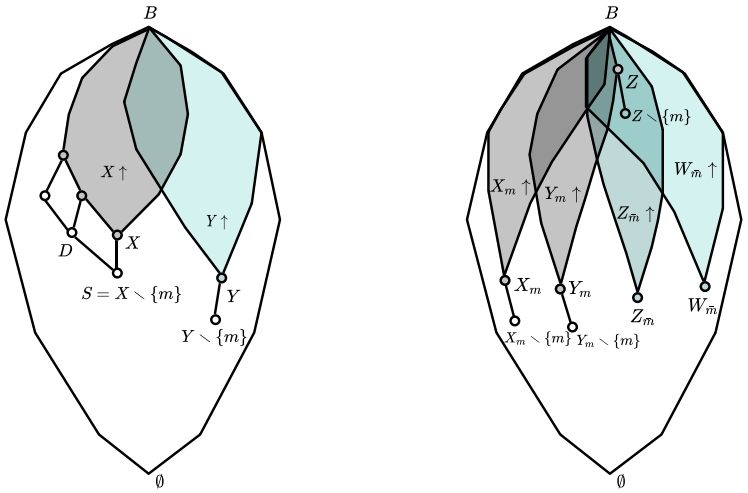

Fig. 1: Schematic line diagrams of $(2^B, \subseteq)$. (a) $X, Y \in \text{mingen}(B)$, $v(X \cup \{m\}) - v(X) = v(Y \cup \{m\}) - v(Y) = 1$, and $v(D \cup X) - v(D) = 1$. (b) X_m, Y_m, $W_{\overline{m}}$ and $Z_{\overline{m}}$ are in $\text{mingen}(B)$ with $m \in X_m, Y_m$ and $m \notin W_{\overline{m}}, Z_{\overline{m}}$. For $Z \in Y_m \uparrow \cap (Z_{\overline{m}} \uparrow \cup W_{\overline{m}} \uparrow)$ we have $v(Z) = 1$ and $v(Z \setminus \{m\}) = 1$ as well.

3. In the last case D and $D \cup \{m\}$ are both generators of B and thus do not contribute to $\varphi_m(A, B)$.

Summing it up, the only contributors are generators of B in the form $D \cup \{m\}$ with $v(D \cup \{m\}) = 1$ and $v(D) = 0$, which finishes the proof.

Corollary 1. *For $m \in X_m \in mingen(B)$, $Y \subseteq B \setminus \{m\}$ with $(Y', Y) \geq (A, B)$ and there is no $Z \subseteq B \setminus \{m\}$ such that $(Y', Y) > (Z', Z) > (A, B)$ we have*

$$\varphi_m(A, B) = \frac{1}{|B|} \sum_{D \in \bigcup [X_m \setminus \{m\}, Y]} \frac{1}{\binom{|B|-1}{|D|}}.$$

Proof. By Theorem 1, since each minimal satisfying set (in terms of set inclusion) is $S \sqcup \{m\} = X_m \in mingen(B)$, then all candidates $D \supseteq S$ for the summation over them lie in the intervals $[X_m \setminus \{m\}, B \setminus \{m\}]$. However, Theorem 1 implies that we also need to eliminate each D in the upper set of a minimal generator without m, $D \in X_{\overline{m}}\uparrow$, since $v(D) = 1$. Thus, we need to consider all maximal sets $Y \subseteq B \setminus \{m\}$ that are not minimal generators of B, i.e. each largest intent of more general concepts for (A, B) that do not contain m fulfils the requirement.

Theorem 2. *Let (A, B) be a concept and $m \in M$, then*

$$\varphi_m(A, B) = \sum_{k=1}^{|B|} \frac{J_k(A)}{k} - \sum_{D \subseteq B \setminus \{m\}} \frac{1}{|D| \binom{|B|-1}{|D|}} v(D) \ . \tag{2}$$

Proof. Let us consider an alternative representation of Shapley value formula as in the original paper [21]. We set $s = |S|$ and get

$$\begin{aligned}
\varphi_m(A, B) &= \sum_{S \subseteq B} \frac{1}{|S| \binom{|B|}{|S|}} (v(S) - v(S \setminus m)) \\
&= \sum_{k=1}^{|B|} \frac{1}{k \binom{|B|}{k}} \sum_{\substack{S \subseteq B: \\ |S|=k}} v(S) - \sum_{S \subseteq B} \frac{(s-1)!(n-s)!}{n!} v(S \setminus \{m\}) \\
&= \sum_{k=1}^{|B|} \frac{J_k(A)}{k} - \sum_{S \subseteq B \setminus \{m\}} \left(\frac{(s-1)!(n-s)!}{n!} + \frac{s!(n-s-1)!}{n!} \right) v(S).
\end{aligned}$$

The identity $\frac{(s-1)!(n-s)!}{n!} + \frac{s!(n-s-1)!}{n!} = \frac{(s-1)!(n-s-1)!}{(n-1)!}$ finishes the proof.

Corollary 2. *If $\{m\} \subseteq X \in \text{mingen}(B)$ and $|\text{mingen}(B)| = 1$, then*

$$\varphi_m(A, B) = \sum_{k=1}^{|B|} \frac{J_k(A)}{k} = \frac{1}{|X|} \ . \tag{3}$$

Another important power index from Choice Theory and Cooperative Game Theory is the so-called Banzhaf power index, which shows how many coalitions that contain a given player are winning, while the same coalitions without that player are not [2]. It is also known that the Banzhaf index is well-correlated with Shapley vector [1]. So, for comparison purposes, let us introduce the weak Banzhaf index of a formal concept where we count all the winning coalitions with the considered player.

Definition 6. *The weak Banzhaf index of an attribute $m \in B$ for the concept (A, B) is defined as follows:*

$$\beta_m(A,B) = \frac{\sum\limits_{D \subseteq B \setminus \{m\}} v_B(D \cup \{m\})}{2^{|B \setminus \{m\}|}} = \frac{|\{D \subseteq B \setminus \{m\} | (D \cup m)' = A\}|}{2^{|B \setminus \{m\}|}} \ .$$

One may note that the weak Banzhaf index of a formal concept is very similar to the extent stability index of a formal concept. In fact, we eliminate contributions of subsets not containing the attribute m from both the numerator and the denominator of the extent stability index and use its computational advantage in our experiments.

5 Machine Experiments

All the experiments in this section have been performed in the interactive environment by means of iPython notebook[8]. To illustrate the differences between the Shapley index and the weak Banzhaf index, we use three contexts of size 3×3 for known elementary scales [6] (Fig. 2) and a positive context describing fruits with binary or scaled attributes from [9] (Fig. 3).

	a	b	c
1	×	×	×
2	×	×	
3	×		

	a	b	c
1	×		
2		×	
3			×

	a	b	c
1		×	×
2	×		×
3	×	×	

Fig. 2: Example contexts for order, nominal, and contranominal scales.

In what follows, we use reduced representations of Shapley vectors, i.e. we hide all the components for attributes outside of a considered concept intent. For 3×3 order scale we get the table below:

Concept	Stability	Shapley value	Banzhaf index
({1}, {a, b, c})	0.5	(0.0, 0.0, 1.0)	(0.5, 0.5, 1.0)
({1, 2}, {a, b})	0.5	(0.0, 1.0)	(0.5, 1.0, 0.0)
({1, 2, 3}, {a})	1.0	(0.0)	(1.0, 0.0, 0.0)

Here, the only non-zero components of the Shapley vectors are given to the attributes that are not contained in the intents of less general concepts, while in the weak Banzhaf indices they have the largest values among other non-zero components.

[8] The iPython script with a related demo can found at `https://github.com/dimachine/ShapStab/`

Table 1: Comparison of Shapley and weak Banzhaf indices for 3×3 nominal scale

Concept	Stability	Shapley value	Banzhaf index
$(\emptyset, \{a, b, c\})$	0.5	(1/3, 1/3, 1/3)	(0.75, 0.75, 0.75)
$(\{3\}, \{c\})$	0.5	(1.0)	(0.0, 0.0, 1.0)
$(\{2\}, \{b\})$	0.5	(1.0)	(0.0, 1.0, 0.0)
$(\{1\}, \{a\})$	0.5	(1.0)	(1.0, 0.0, 0.0)
$(\{1, 2, 3\}, \emptyset)$	1.0		(0.0, 0.0, 0.0)

For the context of nominal scale, the Shapley and weak Banzhaf vectors are equal for all concepts except the top one. As expected, if the intent of a concept of nominal scale consists only of a single attribute, it takes 1 as its importance value. For the top concept, all the values are equally distributed among the attributes, but differ from the values of their counterpart of another index.

Table 2: Comparison of Shapley and weak Banzhaf indices for 3×3 contranominal scale

Concept	Stability	Shapley value	Banzhaf index
$(\emptyset, \{a, b, c\})$	0.125	(1/3, 1/3, 1/3)	(0.25, 0.25, 0.25)
$(\{3\}, \{a, b\})$	0.25	(0.5, 0.5)	(0.5, 0.5, 0.0)
$(\{2\}, \{a, c\})$	0.25	(0.5, 0.5)	(0.5, 0.0, 0.5)
$(\{2, 3\}, \{a\})$	0.5	(1.0)	(1.0, 0.0, 0.0)
$(\{1\}, \{b, c\})$	0.25	(0.5, 0.5)	(0.0, 0.5, 0.5)
$(\{1, 3\}, \{b\})$	0.5	(1.0)	(0.0, 1.0, 0.0)
$(\{1, 2\}, \{c\})$	0.5	(1.0)	(0.0, 0.0, 1.0)
$(\{1, 2, 3\}, \emptyset)$	1.0		(0.0, 0.0, 0.0)

For the context of the contranominal scale, the Shapley and weak Banzhaf vectors are equal except ones for the top concept. If an intent has only one attribute, it takes 1 as its importance value, while it has 0.5 if the intent has exactly two attributes. The values of attributes for the top concept are equally distributed, but the weak Banzhaf index has lower values than those for the top concept of the considered nominal scale since it has only one generator, which coincides with the intent.

As we can see from examples of concepts with fruits like $(\{1, 4\}, \{\bar{f}, s\})$ or $(\{2, 3\}, \{\bar{f}, \bar{s}\})$, attributes that are not contained in minimal generators do not contribute to the Shapley vectors, while their counterparts for weak Banzhaf values do (cf. the importance values of $\bar{f}$).

To experiment with real data, we selected the Zoo dataset (101 examples (animals) and 17 attributes excluding its target attribute) from UCI Machine

Fruits	w	y	g	b	f	$\bar{f}$	s	$\bar{s}$	r	$\bar{r}$
1 apple		×				×	×		×	
2 grapefruit		×				×		×	×	
3 kiwi			×			×		×		×
4 plum				×		×	×			×

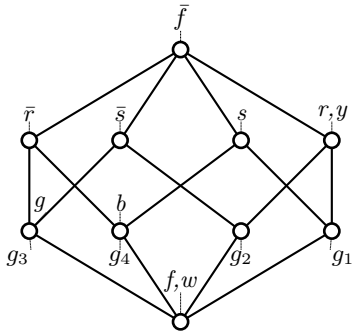

Fig. 3: The context of fruits and the line diagram of its concept lattice.

Table 3: Comparison of Shapley and weak Banzhaf indices for the fruits context

concepts	σ_e	Φ	B
$(\{4\}, \{b, \bar{f}, s, \bar{r}\})$	0.625	(2/3, 0.0, 1/6, 1/6)	(1.0, 0.625, 0.75, 0.75)
$(\{3\}, \{g, \bar{f}, \bar{s}, \bar{r}\})$	0.625	(2/3, 0.0, 1/6, 1/6)	(1.0, 0.625, 0.75, 0.750)
$(\{3,4\}, \{\bar{f}, \bar{r}\})$	0.5	(0.0, 1.0)	(0.5, 1.0)
$(\{2\}, \{y, \bar{f}, \bar{s}, r\})$	0.375	(1/6, 0.0, 2/3, 1/6)	(0.5, 0.375, 0.75, 0.5)
$(\{2,3\}, \{\bar{f}, \bar{s}\})$	0.5	(0.0, 1.0)	(0.5, 1.0)
$(\{1\}, \{y, \bar{f}, s, r\})$	0.375	(1/6, 0.0, 2/3, 1/6)	(0.5, 0.375, 0.75, 0.5)
$(\{1,4\}, \{\bar{f}, s\})$	0.5	(0.0, 1.0)	(0.5, 1.0)
$(\{1,2\}, \{y, \bar{f}, r\})$	0.75	(0.5, 0.0, 0.5)	(1.0, 0.75, 1.0)
$(\{1,2,3,4\}, \{\bar{f}\})$	1	(0.0)	(1.0)

$\sigma_e(\emptyset, \{w, y, g, b, f, \bar{f}, s, \bar{s}, r, \bar{r}\}) = 0.955$
$\Phi = (0.256, 0.069, 0.093, 0.093, 0.260, 0.0, 0.052, 0.052, 0.069, 0.052)$
$B = (1.0, 0.977, 0.984, 0.984, 1.0, 0.955, 0.972, 0.972, 0.976, 0.972)$

Learning Repository[9]. All the attributes are binary except of a single numerical one, the number of legs, which can be treated as categorical and scaled nominally.

This context has 357 concepts in total. We consider the top-3 most stable concepts: c_1, c_2, c_3 along with their extent Stability indices: $\sigma_e(G, \emptyset) = 1$, $\sigma_e(\emptyset, M) = 0.997$, $\sigma_e(A, \{1, 2, 8, 9, 14, 18\}) = 0.625$ respectively, where

$$A = \{11, 16, 20, 21, 23, 33, 37, 41, 43, 56, 57, 58, 59, 71, 78, 79, 83, 87, 95, 100\} .$$

For the top concept, c_1, with empty intent we have zero importance vectors. The attribute names are hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, legs (4), legs (0), legs (2), legs (6), legs (8), legs (5), tail, domestic, and catsize. So, for concept c_2 we can note that

[9] `https://archive.ics.uci.edu/ml/datasets/zoo`

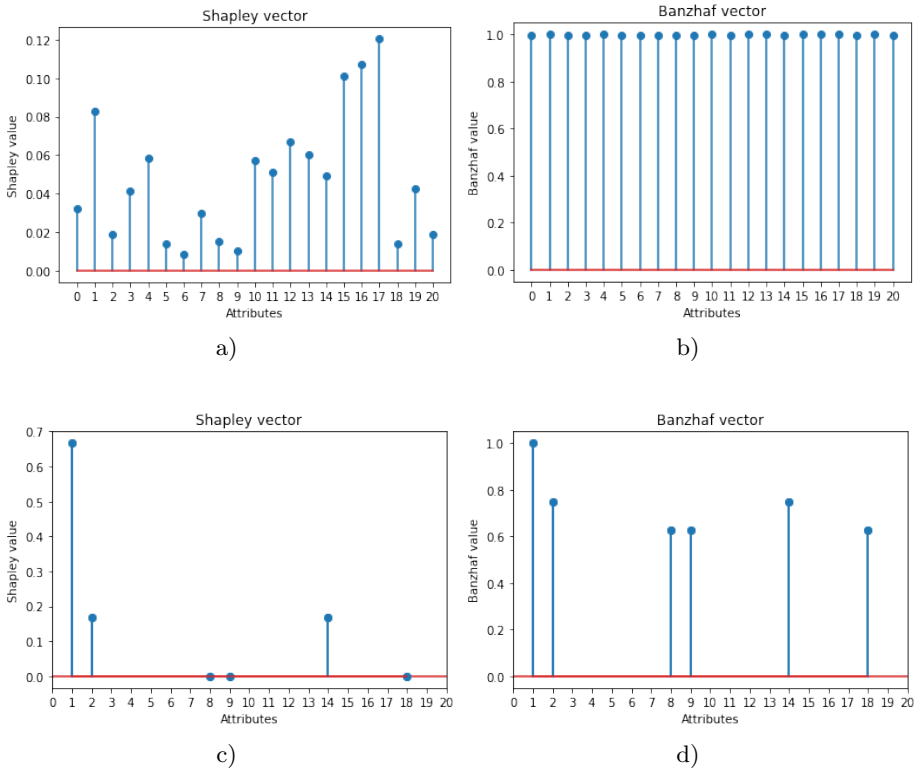

Fig. 4: The Shapley vectors vs. the weak Banzhaf vectors for concepts c_2 (subfigures a) and b)) and c_3 (subfigures c) and d)).

components of the weak Banzhaf vector are nearly ones, while the five largest components of Shapley vector are legs (6), legs (8), legs (5), feathers, and legs (4) (see, Fig. 4). This is the bottom concept with empty extent.

The concepts with empty intent or extent are not very interesting in terms of food for interpretation since they do not describe a discernible class of objects, while concept c_3 looks more attractive with that respect. Even though we have ignored the target attribute with seven classes in the input context, the intent of c_3 describes the class of birds since it consists of the following attributes: feathers, eggs, backbone, breathes, eggs, legs(2) and tail. Among the objects in its intent are chicken, crow, dove, etc. The most important attributes are feathers, eggs and two legs, according to the Shapley vector. They are also the most important ones according to the weak Banzhaf index though it has three remaining attributes with rather high importance values contrary to the Shapley vector.

6 Conclusion

We have introduced the Shapley value and the weak Banzhaf index of a formal concept (or concept intent, more precisely) to rank the attributes of formal concepts based on the associated monotonic Boolean function showing whether a particular set of attributes is a generator of a given concept intent. A similar function is used for the stability indices of formal concepts or JSM-hypotheses [10].

This ranking allows us to order attributes by their importance; in case of the Shapley value, the induced ranking reflects (up to the scaling binomial coefficient) how many times a particular attribute was in a generator of a given concept, while this generator minus the attribute is not a generator of the concept. As we can see from the examples with both toy and real data, these importance values are different from their counterparts in the weak Banzhaf vectors. Thus, they are zeros for all attributes that are not among the attributes of minimal generators.

As we could see from the concept or hypothesis for the bird class in the Zoo dataset, the Shapley value can be used for interpretation purposes selecting the most important attributes of a given hypothesis.

In our further studies we would like to pay special attention to scalability algorithmic issues of this approach (related to reduced contexts) as well as to its comparison with other importance indices [1] from Cooperative Game Theory and Collective Choice with a focus on their interpretability properties for real applications. The detailed discussion on the algorithmic complexity of the related problems will appear in an extended paper version.

Acknowledgements. The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics, and funded by the Russian Academic Excellence Project '5-100'. The first author was also supported by Russian Science Foundation under grant 17-11-01276 at St. Petersburg Department of Steklov Mathematical Institute of Russian Academy of Sciences, Russia. The first author would like to thank Prof. Fuad Aleskerov for the inspirational lectures on Collective Choice.

References

1. Aleskerov, F.T., Habina, E.L., Swartz, D.A.: Binary Relations, Graphs, and Collective Decisions. Fizmatlit (2012)
2. Banzhaf, J.C.: Weighted voting doesn't work: a mathematical analysis. Rutgers Law Review **19**, 317–343 (1965)
3. Faigle, U., Grabisch, M., Jiménez-Losada, A., Ordóñez, M.: Games on concept lattices: Shapley value and core. Discrete Applied Mathematics **198**, 29 – 47 (2016)
4. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. AI Magazine **17**(3), 37–54 (1996)
5. Finn, V.: On Machine-oriented Formalization of Plausible Reasoning in F.Bacon-J.S.Mill Style. Semiotika i Informatika (20), 35–101 (1983), (in Russian)

6. Ganter, B., Wille, R.: Formal Concept Analysis - Mathematical Foundations. Springer (1999)
7. Gionis, A., Mannila, H., Mielikäinen, T., Tsaparas, P.: Assessing data mining results via swap randomization. ACM Transactions on KDD **1**(3), 14 (2007)
8. Kim, B., Koyejo, O., Khanna, R.: Examples are not enough, learn to criticize! Criticism for Interpretability. In: NIPS 2016. pp. 2280–2288 (2016)
9. Kuznetsov, S.O.: Machine Learning and Formal Concept Analysis. In: ICFCA 2004. pp. 287–312 (2004)
10. Kuznetsov, S.O.: On stability of a formal concept. Ann. Math. Artif. Intell. **49**(1-4), 101–115 (2007)
11. Kuznetsov, S.O., Makhalova, T.P.: On interestingness measures of formal concepts. Inf. Sci. **442-443**, 202–219 (2018)
12. Kuznetsov, S.O., Obiedkov, S.A., Roth, C.: Reducing the representation complexity of lattice-based taxonomies. In: ICCS 2007. pp. 241–254 (2007)
13. Kuznetsov, S.: Stability as an estimate of the degree of substantiation of hypotheses derived on the basis of operational similarity. Nauchn. Tekh. Inf. Ser. 2 (12), 217–29 (1991), (in Russian)
14. Kuznetsov, S.: Mathematical aspects of concept analysis. Journal of Mathematical Science **80**(2), 1654–1698 (1996)
15. Lallich, S., Teytaud, O., Prudhomme, E.: Association Rule Interestingness: Measure and Statistical Validation, pp. 251–275. Springer Berlin Heidelberg (2007)
16. Lipton, Z.C.: The mythos of model interpretability. Commun. ACM **61**(10), 36–43 (Sep 2018)
17. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: NIPS 2017. pp. 4765–4774 (2017)
18. Maafa, K., Nourine, L., Radjef, M.S.: Algorithms for computing the Shapley value of cooperative games on lattices. Discr. Appl. Math. **249**, 91 – 105 (2018)
19. Molnar, C.: Interpretable Machine Learning (2019), `https://christophm.github.io/interpretable-ml-book/`
20. Roth, C., Obiedkov, S.A., Kourie, D.G.: Towards concise representation for taxonomies of epistemic communities. In: CLA 2006. pp. 240–255 (2006)
21. Shapley, L.S.: A value for n-person games. Contributions to the Theory of Games **2**(28), 307–317 (1953)
22. Strumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. **41**(3), 647–665 (2014)
23. Tatti, N., Moerchen, F.: Finding robust itemsets under subsampling. In: ICDM 2011. pp. 705–714 (2011)